

Computer Sciences Department

Understanding the World's Worst Spamming Botnet

Tatsuya Mori
Holly Esquivel
Aditya Akella
Akihiro Shimoda
Shigeki Goto

Technical Report #1660

June 2009



Understanding the World’s Worst Spamming Botnet

Tatsuya Mori
NTT Laboratories

Holly Esquivel
UW - Madison

Aditya Akella
UW - Madison

Akihiro Shimoda
Waseda University

Shigeki Goto
Waseda University

ABSTRACT

On November 11, 2008, the primary web hosting company, McColo, for the command and control servers of Srizbi botnet was shutdown by its upstream ISPs. Subsequent reports claimed that the volume of spam dropped significantly everywhere on that very same day. In this work, we aim to understand the world’s worst spamming botnet, Srizbi, and to study the effectiveness of targeting the botnet’s command and control servers, i.e., McColo shutdown, from the viewpoint of Internet edge sites. We conduct an extensive measurement study that consists of e-mail delivery logs and packet traces collected at four vantage points. The total measurement period spans from July 2007 to April 2009, which includes the day of McColo shutdown. We employ passive TCP fingerprinting on the collected packet traces to identify Srizbi bots and spam messages sent from them.

The main contributions of this work are summarized as follows. We first estimate the global scale of Srizbi botnet in a probabilistic way. Next, we quantify the volume of spam sent from Srizbi and the effectiveness of the McColo shutdown from an edge site perspective. Finally, we reveal several findings that are useful in understanding the growth and evolution of spamming botnets. We detail the rise and steady growth of Srizbi botnet, as well as, the version transition of Srizbi after the McColo shutdown.

1. INTRODUCTION

Over the past few years, the volume of e-mail spam has grown significantly to the point it is no longer just a nuisance. Some reports suggest that as much as 90–95% of all e-mail sent or received today is spam [1, 6]. E-mail spam has evolved along many dimensions in recent times, and is employed to conduct numerous subversive and illegal activities such as financial scams, phishing, and propagating malware.

Recently, botnets have been widely used as a *scalable* and *elusive* approach to disseminating spam messages. The bot on the infected host sends out spam messages triggered by instructions from spammers who purchased access to the botnet. Spammers send the instructions from the command and control (C&C) server via IRC channels. Recently, spamming botnets have made the transition from proxy-based spam-

ming to template-based spamming. These new sophisticated user interfaces play a key role in the efficiency of dissemination mechanisms in spamming infrastructures [8]. These improvements have led to an exponential increase in spamming capabilities. For example, “Srizbi” is claimed to be capable of sending 60 billion spam messages per day, which is more than half of the total 100 billion spam messages sent per day on average [9]. According to a more recent report, today’s newest spamming botnet, “Conficker”, consists of more than 10 million infected hosts all over the world and could be capable of sending out 400 billion spam messages per day [2]. These large global-spamming infrastructures have traditionally been hard to stifle.

In late 2008, a bold and drastic action was taken to contain the world’s worst spamming botnet, Srizbi. On November 11, the web hosting service provider, McColo, was shut down by its two upstream ISPs. McColo is known as a so-called “bulletproof hosting” company because it allowed its customers to bypass laws regulating Internet content and services. McColo also allowed these customers to remain online regardless of complaints. The company hosted the C&C servers for major spamming botnets, including Srizbi [3]. Accordingly, as many operators and researchers expected, it is widely reported that the volume of spam dropped from 50 to 75 percent on the very same day [3, 14].

Although recent measurement studies report that spam volume has returned to pre-McColo shutdown levels [14], the temporal but great success of the shutdown indicates that this unprecedented and drastic move was effective. This action allows us to better understand the larger picture of spamming botnets and the way in which they can make transitions, which is crucial to building an effective and sustainable anti-spam solution. As a first step toward this goal, we aim to understand the world’s worst spamming botnet, Srizbi and to study the effectiveness of targeting the botnet’s C&C servers (i.e., McColo shutdown). We also look at the long-term trends of Srizbi to study how the botnet has grown and evolved.

We conduct an extensive analysis of e-mail delivery logs and packet traces collected at four different vantage points across two countries: US and Japan. We also use publicly available packet traces published by MAWI [12]. The four

locations consist of four different types of Internet edge sites, namely, an enterprise network, a campus network, a leaf site of a scientific research network, and an international backbone link used by several research organizations. The total data collection periods span from July 2007 to April 2009. The spam volume changes between the pre- and post-McColo time period can be studied from our data sets.

To detect the spam traffic from Srizbi bots, we leverage a TCP fingerprinting technique, which can identify the operating system of a host based on the TCP/IP stack of the system. As Stern discovered [18], Srizbi uses a dedicated network driver that uses intrinsic TCP/IP parameter settings. Thus, we can extract hosts infected with the Srizbi trojan by tracking their TCP fingerprint signature. In addition to the three signatures presented by Stern et al. [18], we found that the Srizbi botnet has other variants of these signatures.

The primary contributions of this work are:

- We probabilistically estimate the size of Srizbi botnet by correlating data sets that were independently sampled at Internet edge sites.
- We quantify the volume of spam sent from Srizbi and study the effectiveness of the McColo shutdown from the view point of Internet edge sites.
- We reveal several findings that are useful in understanding the spread of spamming botnets; specifically, we note the steady growth of Srizbi and the version transition of Srizbi after the McColo shutdown.

We believe that the collection and analysis of long-term data sets is a promising approach to identifying the upcoming spamming botnets, studying how they are mitigated by actions against them, and building a methodology to stop spamming botnets permanently.

The remainder of this paper is structured as follows: Section 2 presents a description of data sets we use in this work. In section 3, we present our findings on the characteristics and trends of the Srizbi botnet. In section 4 we discuss some related studies and compared them to ours. Finally, section 5 concludes our work.

2. DATA DESCRIPTION

We collected data from four vantage points located at different organizations and countries. The measurement period spans from July 2007 to Apr 2009. The data sets were collected at the University of Wisconsin - Madison, USA; a middle size corporation in Tokyo, Japan; a leaf site of the scientific research network, GEMnet2 [21]; and we also use publicly available data set published by the MAWI WG of the WIDE project [12]. In this work, we call these vantage points UW, CORP, GEM, and MAWI, respectively.

Each vantage point collects one or two primary data sets that are used for this spam analysis. The first set of data collected consists of packet traces of all incoming TCP SYN packets to the SMTP servers. We call this data set tcpdump [19], because that is the name of the network mea-

Table 1: Total measurement periods of data sets.

| | tcpdump | SMTP log |
|------|---|--|
| UW | Feb 9, 2008 – Jul 11, 2008 | Feb 1, 2008 – Apr 30, 2008 |
| CORP | Apr 7, 2008 – Jul 31, 2008 Dec 26, 2008 – Apr 17, 2009 | Apr 7, 2008 – Jul 31, 2008 Jan 1, 2009 – Mar 31, 2009 |
| GEM | – | Aug 1, 2008 – Apr 30, 2009 |
| MAWI | Jul 1, 2007 – Apr 27, 2009 | – |

surement tool used to collect the packet traces. The second set of data contains all e-mail delivery records for each vantage point for all respective e-mail servers. For future reference, we refer to this the data set as SMTP log(s). Table 1 summarizes the measurement period of each data set.

In the following, we describe how each data set is collected and processed for our analysis.

Tcpdump. For UW, and CORP, packet traces are collected on the incoming external links of the networks. For MAWI, we use packet traces which were collected on trans-Pacific line (150-Mbps link) that connects US and Japan, which is utilized by several research organizations. Analyzing packet traces enables us to study all the incoming SMTP connections to the networks. To extract minimal information excluding private information, we filter all the packets other than TCP packets with SYN flag that are destined to the SMTP port. This filtration allows us to employ TCP fingerprinting on packets from e-mail senders while discarding all other private information in the subsequent SMTP transmissions. The IP addresses of MAWI tcpdump traces are anonymized to make the data publicly available. Since these traces have been collected since July 2007, we can study the long-term trends of the spamming botnets.

SMTP logs. For UW, CORP and GEM, SMTP logs were collected on commercial anti-spam appliances. UW and CORP operate greylisting mechanisms on top of their anti-spam appliances. Greylisting is a mechanism that temporarily rejects e-mail messages from a sender which has not previously been seen. Greylisting is effective because if an e-mail is rejected, a spammer will likely not retransmit it since spammers cannot afford the time and resources to retry thousands of bounced messages. By analyzing greylisting logs, the SMTP connections which did not attempt retransmission can be extracted. In this work, we regard these connections as *attempted* spam messages sent to the e-mail servers. That is, if a connection is filtered by greylisting and is not retried later, we regard the connection as a spam message. Note that most spam messages were filtered at the greylisting stage in our data sets. The anti-spam appliances then apply content-based filtering to all messages which *pass* the greylist filtering and spam scores are assigned to them. We adopt conservative thresholds to classify e-mail messages into spam, or ham, based on the score. For example, a spam e-mail must have a spam probability score of greater than 0.95 out of 1.0 in order to be considered spam, while a ham or legitimate e-mail must have a score of smaller than 0.05. In the data sets we analyzed, the majority of messages and connections are classified into spam or ham with the definitions shown

Table 2: Statistics of the SMTP logs for selected months.

| | #spam | #ham | #senders |
|---------------|-------------|------------|-----------|
| Pre-McColo | | | |
| UW Apr 2008 | 101,131,663 | 12,265,296 | 7,473,847 |
| CORP Apr 2008 | 20,107,288 | 545,686 | 2,590,289 |
| GEM Aug 2008 | 95,405 | 1,067 | 68,100 |
| Post-McColo | | | |
| CORP Jan 2009 | 10,886,153 | 723,142 | 1,236,965 |
| GEM Dec 2008 | 65,491 | 2,588 | 36,344 |

above. We note that software-based filtering is error-prone and thus could affect the derived statistics.

Despite the existence of potential errors in the classified messages, the information derived from the scores assigned by the software are sufficient to study the overall characteristics of spamming botnets as we shall see in the next section. Table 2 shows the resulting classification statistics of the logs for selected months. We see that the data sets cover several orders of magnitude. We note that majority of messages seen in all data sets is spam, which is consistent with previous observations [1, 6].

3. ANALYSIS

In this section, we aim to understand the world’s worst spamming botnet, Srizbi and to study the effectiveness of targeting the botnet’s C&C servers, i.e., McColo shutdown. We also look at the long-term trends of Srizbi to better understand how it has been grown and changed. First, we show how we identify hosts infected with the Srizbi trojan (section 3.1). Next, we estimate the size of Srizbi botnet in a probabilistic way (section 3.2). We then quantify the volume of spam sent from the Srizbi botnet, and study the effectiveness of the McColo shutdown from the view point of Internet edge sites (section 3.3). Finally, we reveal the growth of Srizbi botnet and the version transition of Srizbi around the shutdown period (section 3.4).

3.1 Identifying Srizbi

We use a TCP fingerprinting technique to identify Srizbi bots. The signatures are extracted by employing p0f [23] over the collected tcpdump files. From p0f, we are able to analyze specific operating system characteristics about the sending host. The format of the extracted signature is

- [W:T:D:S:O...:Q],

where W stores the information about the window size, T is the initial value of TTL, D is the do-not fragment bit, S is overall SYN packet size, O is the option value and order specification, and Q is a list of miscellaneous information.

Stern [18] carefully studied hosts infected with Srizbi Trojan and found that Srizbi’s TCP/IP driver uses very rare combination of the parameters, which are not used by other operating systems listed in the p0f signatures¹. The following are the three known botnet signatures:

- [24000:128:0:44:M536:.] (Srizbi V1, Ethernet)

¹We manually collected the newer signatures that are not listed on the original p0f signatures, e.g., Windows Vista and Mac OS X 10.5, and found none of them matched to the signatures of Srizbi.

Table 3: Top 5 spam-sending signatures of Srizbi V1 (bold font) and their potential variants for UW (top) and CORP (bottom) in April, 2008.

| signature | #spam | #ham | #senders |
|--------------------------------|------------|-------|-----------|
| UW | | | |
| [24000:128:0:44:M536:.] | 14,495,869 | 2,708 | 260,955 |
| [24000:128:0:44:M1360:.] | 262,077 | 21 | 3,147 |
| [24000:128:0:44:M528:.] | 223,246 | 3 | 2,662 |
| [24000:128:0:44:M1452:.] | 56,589 | 9 | 774 |
| [24000:128:0:44:M1414:.] | 20,504 | 7 | 251 |
| CORP | | | |
| [24000:128:0:44:M536:.] | 7,252,084 | 41 | 1,139,778 |
| [24000:128:0:44:M1360:.] | 126,955 | 0 | 21,329 |
| [24000:128:0:44:M528:.] | 90,518 | 0 | 9,463 |
| [24000:128:0:44:M1452:.] | 30,660 | 0 | 4,025 |
| [24000:128:0:44:M1414:.] | 12,109 | 0 | 2,428 |

- [24000:128:0:44:M528:.] (Srizbi V1, ADSL)
- [6144:255:0:44:M1024:.] (Srizbi V2)

In addition to the above signatures, we found several variants that seemed to be associated with Srizbi. These variants only differ in their MSS values, which reflects the varying window sizes imposed by different types of Internet access links. Table 3 shows the top 5 spam-sending signatures for Srizbi V1 and their variants, sorted by total number of spam messages in the two data sets, UW and CORP.

We notice that the fraction of spam messages sent by the Srizbi signatures are quite high. Variants of the signatures exhibit a similar high fraction of spam messages. Interestingly, the top-5 signatures and their ranking were in common among the two data sets. This indicates that these signatures were stable over time (before McColo shutdown). By looking at the traces collected in later months, e.g., Jan 2009, we can observe Srizbi V2 signatures as well. We omit these results for the brevity.

Finally, as we will detail in section 3.4, the long-term history of these variants signatures exactly agree with the original ones. Thus, we conjecture that these variants are also associated with the original Srizbi. Based on these observations, we add the following signatures as the potential variants of the known Srizbi V1/V2 signatures.

- [24000:128:0:44:M*:.] (Potential Srizbi V1 variants)
- [6144:255:0:44:M*:.] (Potential Srizbi V2 variants)

In the following sections, we leverage these signatures to study the scale of Srizbi botnet, as well as its impact and long-term growth and evolution.

3.2 Estimating Size of Srizbi

Knowing the scale of spamming botnet is useful to estimate the possible worst-case damage caused by a spam flood from a botnet. We leverage a technique proposed by Lawrence and Giles [10] to estimate the size of the Srizbi botnet in a probabilistic way. They estimate the size of indexable web pages on the Internet through the analysis of collected web pages by search engines. To do this, they leverage independently sampled data.

Let $P(X)$ be the probability that a spam bot hits the vantage point X . If we assume that two vantage points A and B

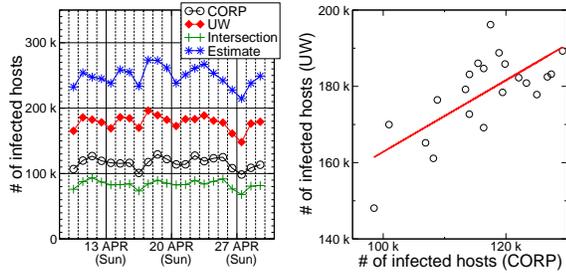


Figure 1: (Left) Estimation of Srizbi botnet; (Right) Scatter plot of active bot sizes observed at CORP and UW. The line indicates the linear regression.

receive spam messages from the Srizbi botnet independently, i.e., a bot selects recipients of spam messages randomly, the probability that a spam bot hits both vantage points is given as $P(A, B) = P(A)P(B)$. Therefore, the total number of hosts infected with Srizbi, $N(\Omega)$, can be estimated as

$$\widehat{N(\Omega)} = \frac{N(A)N(B)}{N(A, B)},$$

where Ω is the entire Internet space and $N(X)$ is the number of spam bots that hit the vantage point X . In this analysis, we use the tcpdump logs of UW and CORP data sets. Because of differing types of these organizations, it is natural to assume that the botnet hits these two sites independently.

When estimating the number of infected hosts it is necessary to take account the reassignment of IP addresses. Zhuang et al. studied the dynamics of IP addresses through the extensive analysis of user login/logout events on Hotmail [24]. They found that about 25% of IP addresses never see IP reassignment in the 7 day log, while a large portion of IP addresses get reassigned almost every day. Based on the above observations, we assume that majority of IP addresses assigned to hosts are stable on a given day; thus, the number of infected hosts seen on a day can be estimated by counting the number of distinct IP addresses seen on that day.

The left panel of Fig. 1 shows the number of IP addresses per day for each data set, their intersections, and the estimated number of active Srizbi-infected hosts per day using the probabilistic model. The analysis uses the data sets collected from 00:00:00, April 9, 2008 to 23:59:59, April 29, 2008 in UTC timezone. We note that the offset of timezones for both sites are corrected. The estimated values range from 210K hosts per day to 275K hosts per day. These numbers agree with the other estimates previously reported in [18] and [8], which claimed that the lower bound of Srizbi botnet size is around 80-130K per day in April 2008 [18] and, the size of Srizbi botnet was around 315K hosts per day in April 2008 [8], respectively.

We also notice that there is clear time synchronization between the number of infected hosts observed at each location. The right panel of Fig. 1 shows a scatter plot of this trend. We see positive correlation between them with a resulting correlation coefficient of 0.715. We conjecture

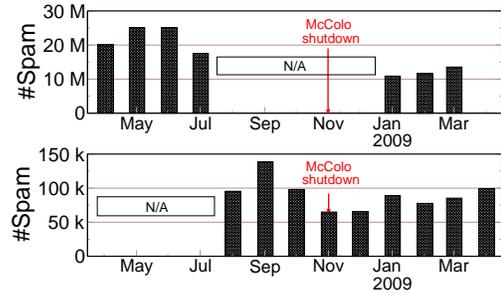


Figure 2: History of spam volumes for CORP (top) and GEM (bottom) data sets.

that the time synchronization reflects the activity of the end-hosts. For instance, the number of hosts decreases every Sunday (in UTC). The way in which a botnet is used, e.g., size of spam campaigns, may also contribute to the global synchronization of botnet activity and the effect of the C&C server shutdown.

3.3 Effectiveness of McColo Shutdown

Here we study how spam volume changed after the McColo shutdown from the view point of Internet edge sites. Figure 2 shows the received spam volumes for CORP and GEM over a period of several months. In both cases, we can see the large reduction in spam volume after the McColo shutdown. In April 2009, the spam volume for GEM data set has almost returned to the pre-McColo level, which agrees with the observation in [14]. On the other hand, the spam volume for the CORP data set has remained at a lower level, i.e., about half of the peak volume, for more than 4 months after the shutdown. According to the network operator of CORP, the level of spam volume is still lower as of early May 2009, which is 6+ months after the shutdown. Thus, the long-term effectiveness of McColo shutdown varies at Internet edge sites.

Next, we study the spam volume decrease from the McColo shutdown relative to the Srizbi botnet. Based on the identification techniques described in Section 3.1, we identify spam messages sent by Srizbi bots. To associate spam messages and TCP fingerprinting, the tcpdump and SMTP logs are correlated together. All the spam messages that appear in the SMTP logs are mapped back to their associated TCP fingerprints found in the tcpdump logs. Table 4 shows the numbers and fractions of spam messages attributed to Srizbi hosts. Prior to the McColo shutdown, a large number of spam messages were sent by Srizbi, but their fractions in total spam volume differ from site to site. For UW, the fraction of Srizbi is around 11–15%. On the other hand, for CORP, the fraction is around 30–45%, which actually has striking impact on the site. We conjecture that the difference in number spam messages reflects the way how the recipients' e-mail addresses are harvested by spammers. Thus, although Srizbi has non-negligible impact on spam volumes of Internet edge sites, its intensity could vary among sites.

Table 4: Breakdown of spam messages sent from Srizbi and other potential end-hosts that have Windows-based TCP fingerprint signatures.

| data set | #total spam | Srizbi (%) | Windows (%) |
|---------------|-------------|------------------|-------------------|
| Pre-McColo | | | |
| UW Feb 2008 | 110,959,667 | 12,602,852 (11%) | 83,333,645 (61%) |
| UW Mar 2008 | 136,572,281 | 17,813,844 (13%) | 101,094,771 (74%) |
| UW Apr 2008 | 101,131,663 | 15,185,849 (15%) | 71,106,454 (70%) |
| CORP Apr 2008 | 20,107,288 | 7,530,864 (37%) | 11,220,937 (56%) |
| CORP May 2008 | 25,079,293 | 10,694,254 (43%) | 13,286,069 (53%) |
| CORP Jun 2008 | 25,088,872 | 11,349,148 (45%) | 12,707,436 (51%) |
| CORP Jul 2008 | 17,562,162 | 5,434,277 (30%) | 10,682,847 (60%) |
| Post-McColo | | | |
| CORP Jan 2009 | 10,886,153 | 607,499 (6%) | 9,487,679 (87%) |
| CORP Feb 2009 | 11,604,039 | 951,914 (8%) | 9,849,693 (85%) |
| CORP Mar 2009 | 13,545,628 | 246,862 (2%) | 12,211,121 (90%) |

We also analyze the source of the remaining spam messages. Table 4 shows the fraction of spam messages sent from Windows hosts. Although the percentage of Windows-based spammers pre-McColo is lower than reported by previous studies, the post-McColo fractions are similar to those seen by Ramachandran and Feamster [16]. We check the IP addresses of these hosts against commercial DNSBL (spamhaus PBL [20]). We found that roughly 90% of IP address space belong to dynamic IP addresses. Thus, although not conclusive, we conjecture that (1) throughout the entire measurement period in Table 4, 72–96% of spam messages are sent from hosts that are likely infected with bots, including Srizbi, and (2) spammers began changing their main spamming infrastructure from Srizbi to other spamming botnets after the McColo shutdown (also see Fig. 3 and 4 for the version transition).

After the McColo shutdown, we see a significant reduction in spam volumes, especially those for Srizbi V1. As a consequent, the total volume of spam messages for the CORP data set is reduced roughly 50% from the pre-McColo level. This indicates that the shutdown has effectively reduced the number of spam messages seen, and hindered a previously prevalent global-scale spam sending infrastructure. Although spam volumes are reviving, continuing action, such as the McColo shutdown, against the source of these messages could be debilitating.

3.4 Long-term trends

To analyze the long-term trends of Srizbi it is desirable to look at several of the data sets. We first use the MAWI data set, which captures packet traces for 15-mins, from 14:00 to 14:15, every day. Although the measured information is sampled in time (sampling rate is 1/96), the data set is useful to track daily trends. Figure 3 shows the number of observed IP addresses with Srizbi V1/V2 signatures and their variants. First, we can clearly observe the rise and death of Srizbi V1, which has been actually terminated since the day of McColo shutdown. In the MAWI data set, the first packet from Srizbi V1 was observed on August 7, 2007. The number of Srizbi hosts observed exceeded a hundred two weeks later and the number kept growing steadily as depicted in the

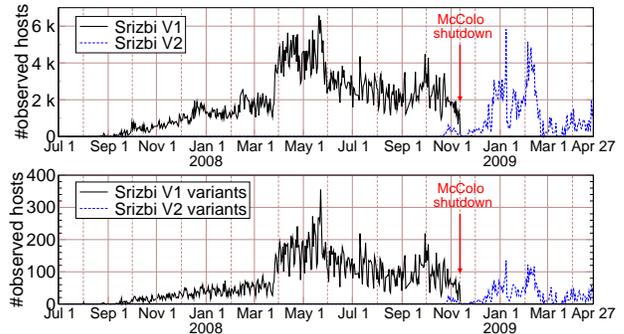


Figure 3: Number of observed hosts infected with Srizbi V1/V2 (top) and their variants (bottom) in the MAWI data set.

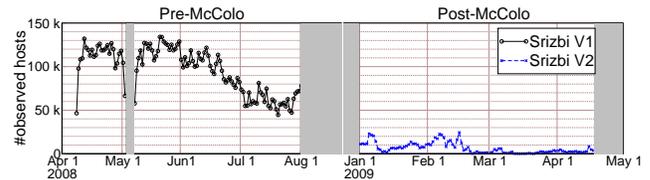


Figure 4: Number of Srizbi V1/V2 hosts observed per day in the CORP data set. Shaded regions are not covered with our measurement.

figure.

We also notice that the spamming history of signature variants are similar to the originals. That is, hosts of V1 and variants emerged around Aug–Sep, 2007 and both were terminated by the McColo shutdown on Nov 11, 2008. Interestingly, Srizbi V2 and its variants have been active since late Oct, 2008. V2 and its variants were soon terminated with McColo shutdown together with V1. Then, activity of Srizbi came back about two weeks later but this time, only V2 and its variants survived. These facts indicate hosts infected with Srizbi and the potential variants were very likely to be sharing the same set of the C&C servers. Thus, we conjecture that newly found variants are associated with Srizbi as we mentioned earlier.

To the best of our knowledge, while many studies such as [14, 18] have reported the resurrection of Srizbi after the McColo shutdown, our study is the first one that presents this version transition around that time period. After the transition, we notice that Srizbi V2 has been less widely spread, compared to V1 before the McColo shutdown. As Stern indicated in [18], the MSRT update in Feb 2009 may have mitigated the spread of V2. This update added a signature for Srizbi to the Malicious Software Removal Tool [13], thus forcing the remaining Srizbi spammers to shift their spamming infrastructures to other spamming botnets.

Finally, Fig. 4 shows the history of hosts infected with Srizbi for the CORP data set. Unfortunately, the data set does not cover the month of McColo shutdown. However, we can observe the transition of Srizbi versions in the time

period before and after the McColo shutdown.

4. RELATED WORK

Botnets have emerged as a major tool for sending spam from end-host machines. To understand the whole picture of spamming botnets, it is crucial to identify hosts infected with spamming bots. Our work leverages TCP fingerprinting to identify hosts infected with Srizbi botnet without their knowledge for analysis. This section first reviews prior studies that identify spamming bots and compare them to ours. We then review several studies that leverage TCP fingerprinting to understand the characteristics of spam senders.

Ways to identify spamming bots have been explored in [5, 15, 17, 18, 22, 24]. Ramachandran et al. [17] develop techniques to identify spamming botnets using passive analysis of DNSBL lookup traffic. The key idea is to find *reconnaissance* lookups from bots. Chiang and Lloyd [5] similarly identified bots, but by monitoring the communication channel between infected hosts and the C&C server of the botnet. Xie et al. [22] developed a framework that outputs high quality regular expressions that can detect messages coming from botnets accurately. Their method successfully identified 7,721 botnet-based spam campaigns, which utilized 340,050 unique IP addresses from a three-month sample of e-mail messages from Hotmail. Also utilizing a Hotmail data set, Zhuang et al. [24] developed a novel technique to extract botnet membership thorough the analysis of e-mail message characteristics. By identifying common characteristics, e-mails can be associated with messages of the same spam campaign.

While the characteristics of spamming botnets have been explored in the previous studies, we build upon this knowledge by exploring a particular spamming botnet in detail and analyzing the effect of the takedown of its C&C servers. Ramachandran et al. [15] monitored DNS queries to the domain hosting the C&C servers of the spamming botnet, Bobax [7], and discovered around 100,000 bot-infected hosts over 46 days. They studied the completeness and responsiveness of popular DNSBLs using the derived IP addresses of the hosts. Stern [18] recently studied the architecture of “Reactor Mailer”, which is a piece of spamware associated with the Srizbi botnet. Through the careful analysis of a large number of hosts infected with Srizbi, they were able to discover the three TCP fingerprinting signatures associated with the botnet. Using the Srizbi signatures, they successfully connected several significant events involving Srizbi botnet. We note that our study reveals several other signature variants of Srizbi V1 and V2, which also contribute a large number of spam messages.

Finally, we review several studies that leverage TCP fingerprint techniques to study the properties of e-mail senders [4, 11, 16]. Ramachandran and Feamster [16] analyzed SMTP traffic delivered to their *spam sinkhole* server and found that approximately 95% of the identified spam-sending hosts were running Windows. Similarly, a study by Li et al. [11] inves-

tigated the operating system information of the spam host machine, using TCP fingerprinting. They found that 74% of the total spam messages were sent from Windows, around 10% were from Linux, about 5% originated from BSD and Solaris machines, and about 11% were from unclassified hosts. Characterizing e-mail based on spamming strategies was proposed by Calais et al. [4]. Their findings suggest a strong correlation between the types of abuse seen and the operating systems from which the abuse originates. Our study reveals similar findings.

5. CONCLUSIONS

The temporal but great success of the McColo shutdown indicates the need for a better understanding spamming botnets as a whole, and the way in which they make transitions is crucial to building effective and sustainable anti-spam solutions. As a first step toward this goal, we studied the world’s worst spamming botnet, Srizbi, and the effectiveness of targeting the C&C servers of the botnet, from the viewpoint of Internet edge sites. We also looked at the long-term trends of Srizbi to study how it has been grown and changed.

First, we estimated the global size of the Srizbi botnet in a probabilistic way. The estimated size ranges from 210k to 275K hosts per day in the April prior to the McColo shutdown. We also found global synchronization within the botnet activity. Knowing the scale and behavior of the spamming botnet is useful to estimate the possible worst-case damage caused by a spam flood from bots. Next, we found that the shutdown was actually effective in reducing the volume of spam at Internet edge sites. For CORP and GEM the spam volume to these sites was reduced by roughly 40-50% and that reduction lasted at least 2-6+ months. We also found the long-term effectiveness of McColo shutdown varies at Internet edge sites. Finally, our analysis of a long-term data set revealed several useful findings in understanding how spammers make transition between spamming botnets. Our analysis revealed the rise and steady growth of Srizbi botnet, and the version transition of Srizbi, triggered by the McColo shutdown.

Our findings suggest targeting a specific set of C&C servers may not be a permanent solution, but it is an effective way to mitigate a significant amount of spam messages at least temporarily. Analyzing the effect of the shutdown is also meaningful to study how spammers make a new transition. Thus, employing new actions against C&C servers of spamming botnet, combined with other methodologies, could eventually reveal in-depth insights into the tricks used by spammers and narrow their options recovery. We believe that keeping an ongoing long-term measurement and analysis is a promising approach for identifying the upcoming spamming botnets, studying how they are mitigated by actions taken against them, and building a methodology to stop spamming botnets permanently. Correlating data sets collected at different layers/locations will play a crucial role in understanding the whole picture of spamming botnets.

6. REFERENCES

- [1] Barracuda Networks Predicts Spam Volumes Beyond 95 Percent in 2009.
http://www.barracudanetworks.com/ns/news_and_events/index.php?nid=322, December 2008.
- [2] Kaspersky Lab analyses new version of Kido (Conficker). <http://www.kaspersky.com/news?id=207575791>, April 2009.
- [3] Brian Krebs. Host of Internet Spam Groups Is Cut Off. http://www.washingtonpost.com/wp-dyn/content/article/2008/11/12/AR2008%111200658_pf.html, November 2008.
- [4] P. Calais, D. Pires, D. Guedes, W. M. Jr., C. Hoepers, and K. Steding-Jessen. A campaign-based characterization of spamming strategies. In *Proc. CEAS 2008: Fifth Conference on Email and Anti-Spam*, 2008.
- [5] K. Chiang and L. Lloyd. A case study of the rustock rootkit and spam bot. In *The First Workshop in Understanding Botnets*, 2007.
- [6] Commtouch. Q3 2007 Email Threats Trend Report. http://www.commtouch.com/downloads/Commtouch_2007_Q3_Email_Threats.pdf.
- [7] Joe Stewart. Bobax Trojan Analysis. <http://www.secureworks.com/research/threats/bobax/>, May 2007.
- [8] Joe Stewart. Top Spam Botnets Exposed. <http://www.secureworks.com/research/threats/topbotnets>, 2008.
- [9] Kelly Jackson. Srizbi Botnet Sending Over 60 Billion Spams a Day. <http://www.darkreading.com/security/encryption/showArticle.jhtml?artic%leID=211201479>, May 2008.
- [10] S. Lawrence and C. L. Giles. Searching the world wide web. *SCIENCE*, 280(5360):98–100, 1998.
- [11] F. Li and M.-H. Hsieh. An empirical study of clustering behavior of spammers and group-based anti-spam strategies. In *Proc. CEAS 2006: Third Conference on Email and Anti-Spam*, 2006.
- [12] MAWI Working Group Traffic Archive. <http://mawi.wide.ad.jp/mawi/>.
- [13] Microsoft. Malicious Software Removal Tool. <http://www.microsoft.com/security/malwareremove/default.aspx>.
- [14] Official Google Enterprise Blog. Spam data and trends: Q1 2009. <http://googleenterprise.blogspot.com/2009/03/spam-data-and-trends-q1-2%009.html>, 2009.
- [15] A. Ramachandran, D. Dagon, and N. Feamster. Can DNS-based blacklists keep up with bots? In *Proc. CEAS 2006: Third Conference on Email and Anti-Spam*, 2006.
- [16] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *Proc. ACM SIGCOMM 2006*, pages 291–302, 2006.
- [17] A. Ramachandran, N. Feamster, and D. Dagon. Revealing botnet membership using DNSBL counter-intelligence. In *2nd Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI)*, 2006.
- [18] H. Stern. The rise and fall of reactor mailer. In *Proc. MIT Spam Conference 2009*, Mar 2009.
- [19] TCPDUMP/LIBPCAP public repository. <http://www.tcpdump.org>.
- [20] The Spamhaus Project. The Policy Block List. <http://www.spamhaus.org/pbl/index.lasso>.
- [21] H. Uose. GEMnet2: NTT's New Network Testbed for Global R&D. In *TRIDENTCOM '05: Proceedings of the First International Conference on Testbeds and Research Infrastructures for the DEvelopment of NeTworks and COMMunities*, pages 232–241, 2005.
- [22] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulthen, and I. Osipkov. Spamming botnets: signatures and characteristics. In *Proc. ACM SIGCOMM 2008*, pages 171–182, 2008.
- [23] M. Zalewski. the new p0f: 2.0.8. <http://lcamtuf.coredump.cx/p0f.shtml>, 2006.
- [24] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, and J. D. Tygar. Characterizing botnets from email spam records. In *Proc. USENIX LEET 2008*, pages 1–9, 2008.